Moses-based official baseline for NEWS 2016

Marta R. Costa-jussà
TALP Research Center
Universitat Politècnica de Catalunya, Barcelona
marta.ruiz@upc.edu

Abstract

Transliteration is the phonetic translation between two different languages. There are many works that approach transliteration using machine translation methods. This paper describes the official baseline system for the NEWS 2016 workshop shared task. This baseline is based on a standard phrase-based machine translation system using Moses. Results are between the range of best and worst from last year's workshops providing a nice starting point for participants this year.

1 Introduction

Transliteration of Name Entities is a useful task for many natural language processing applications such as cross-language information retrieval, information extraction or even machine translation. NEWS workshop has provided for various editions the opportunity to share strategies of transliteration and compare results among different sites. NEWS workshop this year offers training, development and test corpus for 14 language pairs. The final goal of this paper is to offer a baseline system for the NEWS 2016 workshop. Since a general strategy for transliteration has been to use techniques of machine translation, e.g. (Rama and Gali, 2009; David, 2012), we have chosen to use the phrase-based system (Koehn et al., 2003).

The phrase-based machine translation system tries to find the most probable target sentence given the source sentence. The theory behind phrase-based system has evolved from the noisy channel to the log-linear model, which is the one used nowadays. This model combines several feature functions including the translation and language model, the reordering model and the lexical models.

The only requirement to train a phrase-based system is to have a parallel corpus at the level of sentence. In the case of transliteration, we use words as sentences and characters as words. So, for example, parallel sentences to train a transliteration system in English–Hindi is shown in Table 1.

English	Hindi
a a b h a a	अ
a a b h e e r	अ
a a b i d	अ
a a b s h a r	अ

Table 1: Example of English-Hindi Parallel Sentences.

Next experimental section describes the preprocessing of the data and the final corpus statistics for the 14 tasks in the evaluation. We report the parameters used to train the phrase-based system. And finally, we explain the results obtained in terms of several automatic measures. After the experimental section, we include a section of conclusions.

2 Experimental framework

This section describes the corpus statistics that we have used, the parameters of the phrase-based system and the results obtained for each one of the 14 tasks: Arabic-to-English (ArEn), Chinese–English (ChEn, EnCh), English—Thai (EnTh, ThEn), English-to-Persian (EnPe), English-to-Hindi (EnHi), English-to-Tamil (EnTa), English-to-Kannada (EnKa), English-to-Bangla (EnBa), English-to-Korean (EnKo), English-to-Hebrew (EnHe), English-to-Japanese (katakana) (EnJa), and English to Japanese (Kanji) (EnJk).

Proceedings of the Sixth Named Entity Workshop, joint with 54th ACL, pages 88–90,
Berlin, Germany, August 12, 2016.

Languages		Training			Development			Test		
		S	W	V	S	W	V	S	W	V
ArEn	Ar	261.4K	1.5M	38	24,8K	137.9K	38	1.2K	4.5K	6
	En		1.8M	29		181.0K	28		-	-
EnCh/ChEn	Ch	37.7K	119.6K	374	2.7K	9.5K	458	1,0K	2.7K	371
	En		257.7K	26		20K	29	1.0K	6.2K	26
EnTh/ThEn	En	29.6K	210.3K	45	2.0K	14.3K	34	1.2K	8.9K	26
	Th		233.5K	66		15.9K	47	1.2K	10.2K	38
EnPe	En	13.6K	88.0K	26	2.6K	17.0K	26	1.0K	6.3K	26
	Pe		72.3K	43		13.9K	36		-	-
EnHi	En	12.1K	121.7K	44	997	7.1K	26	1,0K	6.3K	27
	Hi		110.9K	83		6.4K	62		-	-
EnTa	En	10.2K	101.9K	42	1.0K	7.2K	29	1.0K	6.3K	27
	Ta		109.7K	63		7.6K	46		-	-
EnKa	En	10.1K	101.0K	42	1.0K	7.2K	30	1.0K	6.3K	27
	Ka		102.6K	75		6.9K	60		-	-
EnBa	En	12.9K	92.7K	30	986	7.0K	27	1,0K	7.0K	27
	Ba		87.8K	62		6.7K	56		-	-
EnKo	En	6.8K	45.4K	28	1.1K	6.1K	26	1.0K	7.5K	28
	Ko		21.4K	714		2.8K	316		-	-
EnHe	En	9.5K	61.3K	32	1.0K	6.4K	26	1,1K	8.1K	28
	He		54.8K	34		5.7K	29		-	-
EnJa	En	31.6K	213.0K	28	1.9K	11.9K	27	1,0K	7,0K	27
	Ja		147.9K	81		8.2K	78		-	-
EnJk	En	23.7K	154.8K	26	3.2K	21.6K	23	1.1K	7.6K	23
	Jk		49.7K	1.6K		6.7K	918		-	-

Table 2: Corpus statistics for training, development and tests sets. S stands for sentences, W for words, and V for vocabulary.

2.1 Data

Table 2 details the corpus statistics for all 14 tasks including training, development and test sets. Preprocessing has been limited to separate characters by a blank space.

2.2 System Description

The phrase-based system was built using Moses (Koehn et al., 2007), version 15th April 2016 from github, with standard parameters, including: grow-final-diag for alignment; Good-Turing smoothing of the relative frequencies; 3-gram language modeling using Kneser-Ney discounting and training with SRILM (Stolcke, 2002); and lexicalized reordering, which includes 6 feature functions. Optimization was done using the MERT algorithm and MBR option for decoding. It is important to note that the same system was used for the 14 tasks without any change or modification.

2.3 Results

Official results are reported in Table 3. In most tasks, results were in the middle of the ranking. Best ranking results were obtained in English-to-Japanese (Kanji) and Arabic-to-English (no merit this one, because the baseline was the only participant). Worst ranking results were for English–Thai, English-to-Tamil, English-to-Hebrew, English-to-Korean, English-to-Japanese (Katakana).

3 Conclusions

This phrase-based system based on standard Moses has been offered to the NEWS organizers to provide a reasonable baseline system for the competition. Also, it helps the participants to know the quality level of their systems compared to state-of-the-art transliteration when faced as a translation challenge.

In the next edition, we hope to provide an en-

Task	ACC	F-Score	MRR	MAP
ArEn	0.4809	0.9127	0.4809	0.1275
EnCh	0.1934	0.5850	0.1934	0.1830
ChEn	0.0098	0.6459	0.0981	0.0953
EnTh	0.0679	0.7069	0.0679	0.0679
ThEn	0.0914	0.7396	0.0914	0,0914
EnPe	0.4817	0.9060	0.4817	0.4482
EnHi	0.2700	0.7992	0.2700	0.2624
EnTa	0.2580	0.8116	0.2580	0.2572
EnKa	0.1960	0.7832	0.1960	0.1955
EnBa	0.2870	0,8359	0.2870	0.2837
EnHe	0.1090	0.7714	0.1090	0.1077
EnKo	0.2130	0.6177	0.2180	0.2176
EnJa	0.2091	0.7047	0.2091	0.2059
EnJk	0.461	0.6517	0.4611	0.2967

Table 3: Official NEWS 2016 Results.

hanced baseline system by tuning some parameters from the Moses system, and possibly competing in the shared task with some related approach to character-aware neural machine translation system (Costa-jussà and Fonollosa, 2016).

Acknowledgements

This work is supported by the 7th Framework Program of the European Commission through the International Outgoing Fellowship Marie Curie Action (IMTraP-2011-29951) and also by the Spanish Ministerio de Economía y Competitividad and European Regional Developmend Fund, contract TEC2015-69266-P (MINECO/FEDER, UE). Author also wants to specially thank Dr. Rafael E. Banchs for motivating these experiments.

References

Marta R. Costa-jussà and José A. R. Fonollosa. 2016. Character-based neural machine translation. In Proceedings of the ACL.

Chris Irwin David. 2012. Tajik-farsi persian transliteration using statistical machine translation. In Proceedings of the LREC.

Philipp Koehn, Franz Joseph Och, and Daniel Marcu. 2003. Statistical Phrase-Based Translation. In Proceedings of the ACL.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicolas Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open Source Toolkit for Statistical Machine Translation. In Proc. of the ACL, pages 177–180.

Taraka Rama and Karthik Gali. 2009. Modeling machine transliteration as a phrase based statistical machine translation problem. In Proceedings of the Named Entities Workshop: Shared Task on Transliteration, pages 124–127.

Andreas Stolcke. 2002. Srilm-an extensible language modeling toolkit. In Proceedings International Conference on Spoken Language Processing, pages 257–286, November.